DOCUMENT RESUME

ED 059 236 TM 001 045

AUTHOR	Carroll, John B.
TITLE	Measurement Properties of Subjective Magnitude Estimates of Word Frequency.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	RB-71-45
PUB DATE	Jul 71
NOTE	19p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Adults; Correlation; Discriminant Analysis; *Individual Differences; Lexicography; *Prediction; Probability Theory; Rating Scales; *Reliability; Transformations (Mathematics); *Word Frequency; Word Recognition
IDENTIFIERS	SME; *Subjective Magnitude Estimation

ABSTRACT

The subjective magnitude estimation (SME) procedure was used to obtain estimates of relative word frequency from two adult groups (15 lexicographers, 13 other adults) for 60 words ranging widely in objective frequency. Lexicographers rendered more reliable estimates, and their averaged data correlated more highly with objective log frequency than those of the second group. The objective frequency of the first stimulus considered in the SME task is not related to a subject's overall accuracy in predicting objective frequency, but accuracy is related to the subject's tendency to perceive frequency ratios as relatively large. Subjective estimates measure something slightly different from what is indexed by currently available objective counts, and may be more valid measures of true word probability. (Author/CK)

ED 059236

RB-71-45

RESEARCH BULLETIN

MEASUREMENT PROPERTIES OF SUBJECTIVE MAGNITUDE ESTIMATES OF WORD FREQUENCY

John B. Carroll

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

This Bulletin is a draft for interoffice circulation. Corrections and suggestions for revision are solicited. The Bulletin should not be cited as a reference without the specific permission of the author. It is automatically superseded upon formal publication of the material.

Educational Testing Service
Princeton, New Jersey
July 1971

TM 001 045

Measurement Properties of Subjective Magnitude Estimates of Word Frequency

John B. Carroll

Educational Testing Service

Abstract

Stevens' subjective magnitude estimation (SME) method was used in obtaining estimates of relative word frequency from two adult groups (15 lexicographers, 13 other adults) for 60 words ranging widely in objective frequency. Lexicographers rendered more reliable estimates, and their averaged data correlated more highly (.970) with objective log frequency than those of the second group (.923). The objective frequency of the first stimulus considered in the SME task is not related to an S's overall accuracy in predicting objective frequency, but accuracy is related to the S's tendency to perceive frequency ratios as relatively large. Subjective estimates measure something slightly different from what is indexed by currently available objective counts, and may be more valid measures of true word probability.

Measurement Properties of Subjective Magnitude Estimates of Word Frequency[1]

John B. Carroll

Educational Testing Service

The subjective magnitude estimation (SME) procedure developed by Stevens (1956; 1958, p. 193) has been successfully applied by Shapiro (1969) to the scaling of word frequency; Shapiro found correlations of from .920 to .958 between subjective estimates and objective data on word frequency when both variables were in logarithmic form. The purpose of the present study was to examine the reliability of such subjective estimates as a function of the number and characteristics of the raters. A further purpose was to investigate individual differences in the accuracy of subjective estimates (i.e., in the correlations of Ss' judgments with objective data) and in the exponents implied by the power-law function. Consideration was given to whether these differences are a function of the objective frequency of the first word considered in the SME task, in view of the fact that the response assigned to that word tends to establish the characteristics of the arbitrary scale used by the S in making his subsequent responses. There was also interest in whether evidence could be obtained as to whether subjective estimates are more valid indices of true word probability than currently available objective word-frequency counts.

Method

Subjects. Two samples of subjects were used. One (N = 15) was a group of individuals who as a consequence of serving on the editorial staff of the American Heritage Dictionary (Morris, 1969) had had extensive

experience in lexicography. These lexicographers might be expected to have particularly accurate perceptions of word frequency. The other sample (N = 13) was composed of teachers, housewives, and other adults, all with college educations, who volunteered for the study. Ss were not paid for participation; they were allowed to perform the task at their own convenience, but were cautioned not to refer to any objective word-count data in making their judgments.

Materials. The stimuli were 60 words that had been used by Shapiro (1969); these are listed in descending order of objective frequency in the first column of Table 1. These words had been selected by Shapiro to represent a wide range of frequency according to two large frequency counts (Thorndike & Lorge, 1944; Kučera & Francis, 1967). In the second column of Table 1, the raw frequencies assembled by Shapiro are converted to SFI (Standard Frequency Index) values according to the formula proposed by Carroll (in press):

$$SFI = 10\ (\log_{10} f + 4\ ),$$

where f = frequency per million. The words were printed in three columns on computer-output paper in the same random order that had been used by Shapiro (1967, p. 194). Associated with each word, however, was a computer-generated number to indicate the order in which the words were to be considered in the judgment process; the rating form given to each S had a different random order of these numbers, ranging from 1 through 60.

Procedure. With slight modifications, Shapiro's instructions for the subjective magnitude estimation procedure were used. Ss were told that their task was to "tell with numbers" how frequently the words occur in written English. "Give one of the words any number that seems appropriate to you. Then give numbers to the other words so that these

numbers give your own impression of the relative frequency of these words in written English. Thus, if a word seems 20 times as frequent as another, give it a number 20 times as large. If, however, it seems only half as frequent, give it a number only one-half as large.... It is best to use a first number that is fairly large so that you have room to move both up and down. You may use fractions, decimals, or whole numbers, but not negative numbers. You will not be timed. Work at your own speed. There are no right or wrong answers. We are simply interested in your own judgments." After a brief practice in scaling the words patron, salad, dictum, and mother, in that order, the S was told: "On the accompanying sheet, the 60 words are listed in random order, but each word is tagged with a number indicating the order in which you are to take up the words. Please follow this order, starting with (1), (2), etc., until you reach (60). It should not be too difficult to skip around to find the order numbers." Ss wrote their responses on blanks placed to the right of each word, and could at any time inspect or change the responses they had made.

Results

For each S, the following statistical operations were performed: (1) The numbers given for each word were converted to logarithms to base 10; (2) the mean and S.D. of these log values were computed, and the log values were then converted to z-scores so that comparable values for the words could be obtained over the different Ss; (3) the correlation between the log values and the SFI values was found, along with the corresponding regression equations.

For each sample of Ss, the following statistics were then computed: (1) the mean and S.D. of the mean log values and of the mean z-scores;

(2) for each word, the mean log value, the mean z-score value, and the S.D. of the z-scores over the Ss; (3) over the 60 words, the correlations between the objective SFI values and (a) the mean log values and (b) the mean z-scores, with the corresponding regression equations; (4) reliability coefficients for individual and group-average z-scores by the methods of Ebel (1951), indicating degree of inter-rater agreement.

The mean log values, the mean z-score values, and the S.D.'s of z-scores are given for each word and for each word in Table 1. Data are also given for combined samples.

- - - - - - - - - - - - - - -
Insert Table 1 about here
- - - - - - - - - - - - - - -

Sample 1 (N = 15 lexicographers) yielded significantly more reliable estimates than Sample 2 (N = 13 other adults) as determined by several criteria. As shown in Table 2, the respective reliability coefficients

- - - - - - - - - - - - - - -
Insert Table 2 about here
- - - - - - - - - - - - - - -

for individual ratings in z-score form were .869 and .746, significantly different by Fisher's test (t = 2.00, df = 114, $p < .05$, each correlation being based on n = 60 words). Using the Spearman-Brown prophecy formula to boost the obtained reliability for the average z-score ratings of Sample 2 (.974) to make it comparable with that of Sample 1, we obtain .978, which is significantly different from the obtained reliability of average ratings for Sample 1, .990 (t = 2.16, $p < .05$). The mean square errors for the two groups are also significantly different, .134 and .259 respectively ($F_{708,826} = 1.94$, $p < .001$). This difference is also reflected in the fact that the median S.D. of z-scores over the 60 words is significantly lower in Sample 1, .327, than in Sample 2, .442 (by Wilcoxon's test for nonindependent samples, $z = 5.04$, $p < .001$).

There is, however, a tendency for the individual z-scores to be increasingly unreliable as the SFI values decrease, i.e., as the stimulus words become less and less frequent in terms of objective data. Table 3 shows average S.D.'s of z-scores by groups of words according to SFI values. As seen there, the S.D.'s of z-scores yielded by lexicographers tend to be in the neighborhood of .32 for all words except those of lowest frequency, average SFI = 37.4, where their average S.D. is .482 ($F_{5,54} = 9.6$, $p < .001$). In Sample 2, however, the S.D.'s are generally higher than those of the lexicographers, and these S.D.'s increase markedly as the SFI values decrease ($F_{5,54} = 7.12$, $p < .001$). There was a small yet significant correlation ($r = .272$, $df = 57$, $p < .05$) between the S.D.'s of z-scores in the two samples.

- - - - - - - - - - - - - - -
Insert Table 3 about here
- - - - - - - - - - - - - - -

Using the reliability data obtained for the groups, we can estimate standard errors of measurement for the individual z-scores and for the averaged z-scores. These standard errors of measurement can also be translated in terms of SFI units. The relevant results are given in Table 3 for the two groups, and for the two groups combined. Some explanation is in order. In the case of the z-scores derived for an individual's judgment for a particular word, the reliabilities and mean square errors are those computed by Ebel's (1951) method. The standard errors of measurement in z-score units are the square roots of the mean square error values. To translate these into terms of SFI units, it is assumed that the correlation between z-scores and SFI values is unity; since the z-scores for any given individual have an S.D. of unity, we multiply the standard errors of measurement by the S.D. of the SFI values, which is 13.896.

In the case of the averaged z-scores, the reliabilities are those obtained by the Ebel method. The standard errors of measurement are then obtained by the formula $S.E._{measmt} = \sigma_z\sqrt{1 - r}$, where σ_z is the S.D. of the obtained averaged z-scores for the words. Translation of these standard errors of measurement into SFI units again assumes perfect correlation between the SFI values and the averaged z-scores; however, since the S.D.'s of these averaged z-scores are not unity, we multiply the standard errors of measurement by the ratio of the respective standard deviations, i.e., $13.896/\sigma_z$.

In all cases the 95% confidence limits are obtained by multiplying the standard errors of measurement by $\pm$ 1.960 to include 95% of the area in the normal curve.

These results imply that if we regard an individual's judgment of the frequency of a particular word as an independently valid measure, and translate it into SFI units by the formula: $SFI = 56.3 + 13.9z$, the resulting value will be within $\pm$ 9.95 or $\pm$ 13.85 of the "true" value 95% of the time, depending on whether he is a lexicographer or a nonlexicographer, respectively. Roughly speaking, this means that an individual can usually judge word-frequency within one order of magnitude.

Similarly, if we regard the average z-score value given to a word by a group such as our samples as an independently valid measure of frequency, and translate it into SFI units by the formula $SFI = 56.3 + 13.9z/\sigma_z$, the resulting value will be within $\pm$ 2.72 or $\pm$ 4.35 of the "true" value 95% of the time, for groups of lexicographers and nonlexicographers, respectively. This degree of accuracy is attained when data are averaged over about 14 raters, but even higher reliability coefficients and tighter confidence limits are obtained by using data averaged over all 28 raters, as shown in Table 2.

It is of interest, of course, to consider to what degree the individual and averaged estimates (in terms of both log values and z-score values) are actually correlated with SFI values. First consider the correlations for z-scores derived for particular individuals. (In this case the correlations with SFI are the same for log values and for z-scores since the latter are linear transforms of the former.) Lexicographers' subjective estimates tended to be more highly correlated with objective frequency ratings than those of the other sample. For them, the correlations ranged from .781 to .972 and had a median of .921, whereas for the other adults, the correlations ranged from .657 to .918 and had a median of .827; these medians are significantly different by the Mann-Whitney test ($z = 3.25$, $p < .01$).

The two procedures of averaging individual ratings gave only slightly different results. When individual log values were averaged, the correlations with SFI values were .974 and .929 in the two samples, respectively; these are significantly different by Fisher's test ($t = 2.76$, $df = 114$, $p < .01$). The correlations for averaged z-scores were .970 and .923, respectively, again significantly different ($t = 2.53$, $p < .02$). For data averaged over the combined samples, the correlations were .966 for log values and .959 for averaged z-scores. All these correlations, of course, are highly significantly different from zero, being based on $n = 60$. The two sets of scores had an intercorrelation of .957 both for average log scores and for average z-scores. This result will be remarked on in the discussion.

A feature of the subjective magnitude estimation procedure is that it enables determination of a "psychophysical" relation between objective experience and subjective judgment. If we assume Stevens' power law to apply in the case of word frequency judgments, the relationship should

be of form

$$j = af^m ,$$

or, taking logarithms,

$$\log j = \log a + m \log f,$$

where j = judged frequency and f = objective frequency, and where j and f both have arbitrary origins and units. There is interest in the magnitude of the exponent m, since it expresses how rapidly subjective frequency increases as a function of experienced frequency. The value of the constant a is, however, dependent solely on the origins of the scales in which the measurements are expressed. The present data yield some evidence on the values of these constants.

First, consider the case where the judgment is assumed to be perfectly correlated with the experienced frequency. The standard deviation of the SFI values of the stimuli used in this experiment was 13.896, but since the SFI values represent logarithmic values multiplied by a factor of ten, we may take the standard deviation of the log frequencies as being equal to 1.390. The standard deviations of the log values assigned by each individual over the words were obtained. The ratios of these standard deviations to the S.D. of the stimulus logs may then be regarded as representing values of m. For Sample 1 (lexicographers), these ratios ranged from .410 to 2.041, with a median of .917; when the ratio is based on the S.D. of the averaged log values, it becomes .854. For Sample 2, the individual ratios ranged from .536 to 1.48 with a median at .654; on the basis of the averaged log values the ratio is .685. Thus, on the assumption of a perfect correlation between judgment and experienced frequency our samples tend slightly to underestimate ratios of frequencies, since the ideal logarithmic ratio would be 1.00. Nevertheless, there are wide individual differences, attributable partly to differences in people's

perceptions of ratios of word frequencies and partly, perhaps, to differences in their habits of handling numbers.

When the regressions of log values on SFI values (divided by 10) are considered, the ratios are obtained directly from the slope coefficients. Regression equations were obtained for each individual and for averaged data. For Sample 1, the ratios ranged from .358 to 1.946 with a median at .868; on the basis of averaged data, the ratio was .832. For Sample 2, the ratios ranged from .379 to 1.227 with a median of .507; for averaged data, the coefficient was .581.

There is no particular interest in the values of (log a) since these values merely indicate the log values that would presumably (by extrapolation) have been assigned to a word having a value of SFI = 0. It happens that SFI is scaled so that it takes the value of 0 when the word probability is 10^{-10}, or 1 in 10 billion. Thus, these values of (log a) are a function of the general sizes of numbers employed by the Ss. However, it is of some interest to note that the values of averaged z-scores corresponding to SFI = 0 were -3.677 for Sample 1 and -3.270 for Sample 2, based on the regression equations of averaged z-scores on SFI values. These numbers, of course, are well outside the range of averaged z-values actually obtained, a result that is reasonable in view of the fact that there were no stimulus words as rare as one with a probability of 10^{-10}.

Requiring each subject to consider the words in a different random order made it possible to investigate whether the relative frequency of the first word considered made for differences in the accuracy or other aspects of the subjective judgments. It was thought that the objective relative frequency of that word might influence the number that the subject would assign to the word, and in that way also influence the characteristics of the scale used for the subsequent

judgments, and indirectly, the accuracy of those judgments. Over the combined samples ($N = 28$), there was a correlation of .458 ($p < .05$) between the SFI of the first stimulus word and the log value of the number assigned to it; the corresponding correlation with the z-score derived from that log value was .922 ($p < .001$). This means that Ss tend to choose a "first number" or anchor value in accordance with its perceived relative frequency. (This is not always the case; one S assigned "1" to the first word, which was other, with SFI = 72.2, and still managed to render highly accurate judgments by using mainly fractional numbers for the remaining words.) The log of the "first number" chosen by the Ss was significantly correlated ($r = .655$, $p < .01$) with the mean log response to all words, a result that suggests that this number tends to establish the scale on which the remaining words are judged. Nevertheless, the log magnitude of that number was not significantly correlated ($r = .122$, $p > .05$, with no evidence of significant curvilinearity) with an S's accuracy in estimating objective frequencies, when accuracy is indexed by the correlation between the log of his subjective judgment and SFI. Nor was the average log value of an S's responses over all stimuli significantly related to his accuracy ($r = -.120$). The standard deviation of the log values of his responses, however, was significantly correlated with his accuracy ($r = .433$, $p < .05$).

We may conclude from these data that although the objective relative frequency of the first word considered has some influence on the scale used for the subsequent words, it does not influence the accuracy of the judgments. It is immaterial, in the SME task, whether the first word judged is a high, medium, or low frequency word. However, there is a tendency for Ss who perceive frequency ratios as relatively

large to render more accurate estimates than Ss who perceive frequency ratios as relatively small.

Discussion

The data give evidence that the subjective estimates are consistently deviant from the objective frequency values. The correlation between the two sets of averaged z-scores was .957; the two sets correlated with SFI values to the extent of .970 and .929, respectively. If we partial out the SFI variable, the inter-set correlation stands at .657 ($df = 57$, $p < .001$). This result suggests that the subjective estimates of frequency are not measuring exactly the same thing as the objective data. There arises the question of which of these measures is more valid with respect to "true" word probability. It can be argued that the subjective estimates are more valid, on the grounds that objective frequency counts such as the Thorndike-Lorge count are subject to biases of various kinds in sampling, in establishing units, etc., and that human observers are better able to discount such biases. Furthermore, the samples used in objective frequency counts, seldom more than a few million tokens, are small in comparison to the number of tokens experienced by the human observer over his lifetime. Objective frequency counts are particularly subject to bias in the case of words of low probability. If the true probability of a word is less than $1/N$, where N is the number of tokens in a sample, and if that word appears at all in the sample, by chance, its observed probability is at least $1/N$, a value that is biased upward. For example, a word with a true probability of 1 in a million that happens to occur once in a sample of 1000 words will have an observed probability of 1 in a thousand. In the present data, it can be observed by plotting z-scores against SFI

values that particularly in the case of low frequency words, the z-scores are relatively lower than the SFI values in relation to the line of equivalence (the line of equal standard deviation scores for the two variables). Thus, the argument that subjective estimates are more valid than objective data tends to be stronger in the case of low frequency words.

Ultimately, the question of the validity of subjective estimates can probably be settled only by comparing them with data obtained from much larger and more refined frequency counts than are currently available. The high degree of precision that can be attained by averaging SME data over relatively small numbers of raters recommends this technique as likely to be useful in assessing the frequency attribute of stimuli in verbal learning experiments. By their very nature, subjective estimates reflect perceived frequency and hence have more immediate psychological relevance than word count data.

REFERENCES

Carroll, J. B. An alternative to Juilland's usage coefficient for lexical frequencies, and a proposal for a Standard Frequency Index (SFI). Computer Studies in the Humanities and Verbal Behavior, in press.

Ebel, R. L. Estimation of the reliability of ratings. Psychometrika, 1951, 16, 407-424.

Kučera, H., & Francis, W. N. Computational analysis of present-day American English. Providence, R.I.: Brown University Press, 1967.

Morris, W. (Ed.) The American Heritage dictionary of the English language. New York and Boston: American Heritage Pub. Co. and Houghton Mifflin, 1969.

Shapiro, B. J. The subjective scaling of relative word frequency. (Doctoral dissertation, Harvard University) Cambridge, Mass.: University Microfilms, 1967. No. 67-12,549.

Shapiro, B. J. The subjective estimation of relative word frequency. Journal of Verbal Learning and Verbal Behavior, 1969, 8, 248-251.

Stevens, S. S. The direct estimation of sensory magnitudes--loudness. American Journal of Psychology, 1956, 69, 1-25.

Stevens, S. S. Problems and methods of psychophysics. Psychological Bulletin, 1958, 55, 177-196.

Thorndike, E. L., & Lorge, I. The teacher's word book of 30,000 words. New York: Teachers College Press, 1944.

FOOTNOTE

[1]This work was supported by the National Institute of Child Health and Human Development, under Research Grant 1 P01 HD01762. I thank Mr. Barry Richman, of the American Heritage Publishing Co., Inc., for securing the cooperation of both groups of subjects and for arranging for the collection of the data. I also thank Miss Barbara Witten for her assistance in the computations and data analysis.

Table 1

Mean Log Values, Mean z-Scores, and Standard Deviations of z-Scores, for Sample 1, Sample 2, and Combined Samples, for Subjective Magnitude Estimates of 60 Words, with SFI Values of the Words

		Sample 1 (N = 15)			Sample 2 (N = 13)			Combined Samples (N = 28)		
Word	SFI	Mean Log	Mean z	S.D. z	Mean Log	Mean z	S.D. z	Mean Log	Mean z	S.D. z
the	88.4	4.95	2.01	.475	4.57	1.46	.446	4.77	1.75	.537
of	85.6	4.59	1.68	.318	4.30	1.23	.345	4.46	1.47	.400
and	84.5	4.69	1.79	.327	4.48	1.40	.300	4.59	1.61	.370
that	80.2	4.34	1.47	.341	4.20	1.15	.246	4.28	1.32	.340
as	78.5	4.18	1.39	.279	4.14	1.09	.400	4.16	1.25	.372
by	77.2	4.16	1.37	.328	4.07	1.09	.400	4.12	1.24	.389
you	75.1	4.31	1.44	.304	4.20	1.06	.509	4.26	1.26	.454
when	73.6	3.95	1.16	.340	4.07	.97	.557	4.01	1.07	.464
other	72.2	3.73	1.02	.400	3.90	.82	402	3.81	.93	.413
after	70.2	3.68	.98	.288	3.96	.89	.289	3.81	.94	.292
each	69.4	3.70	.98	.264	3.98	.92	.377	3.83	.95	.323
few	67.7	3.23	.69	.330	3.54	.63	.430	3.37	.66	.381
again	67.6	3.57	.92	.254	3.82	.80	.206	3.69	.86	.240
night	66.1	3.07	.52	.252	3.61	.57	.381	3.32	.54	.319
next	65.9	3.35	.74	.216	3.91	.83	.278	3.61	.78	.251
early	65.6	3.09	.56	.341	3.35	.44	.372	3.21	.50	.361
half	64.3	3.26	.71	.338	3.45	.47	.247	3.35	.60	.322
result	63.8	2.97	.40	.327	3.41	.45	.241	3.17	.42	.291
music	63.3	2.82	.32	.362	3.43	.37	.523	3.10	.34	.445
final	61.9	2.82	.32	.410	3.35	.38	.248	3.07	.35	.346
list	61.2	2.65	.21	.295	3.63	.53	.444	3.10	.36	.404
couple	60.8	2.82	.42	.331	3.55	.51	.309	3.16	.46	.324
price	60.3	2.77	.31	.331	3.65	.59	.341	3.18	.44	.364
actual	60.0	2.59	.20	.418	3.35	.45	.473	2.94	.32	.462
base	59.5	2.66	.20	.227	3.34	.36	.627	2.98	.27	.465
spread	59.1	2.61	.19	.334	3.24	.22	.422	2.90	.20	.378
address	58.8	2.63	.21	.275	3.37	.31	.333	2.97	.26	.307
scale	57.7	2.38	.03	.251	2.99	.11	.303	2.66	.07	.279
suit	56.7	2.69	.23	.249	3.35	.40	.321	3.00	.31	.297
humor	56.6	2.54	.11	.242	3.18	.15	.396	2.84	.13	.323
swift	56.3	2.25	- .09	.299	2.97	.14	.450	2.58	.02	.394
victim	55.6	2.26	- .15	.348	3.09	.22	.550	2.65	.02	.489
anchor	54.1	1.92	- .33	.215	2.67	- .15	.675	2.27	- .25	.494
convert	54.0	2.13	- .22	.241	3.02	.00	.458	2.54	- .12	.375
charter	53.0	1.86	- .40	.177	2.64	- .22	.538	2.22	- .32	.399
stride	52.6	1.97	- .31	.226	2.39	- .44	.493	2.16	- .37	.380
switch	51.8	2.49	.06	.204	2.98	.07	.336	2.72	.06	.273
volcano	51.5	1.81	- .44	.281	2.37	- .56	.459	2.07	- .50	.379
heritage	49.0	1.80	- .46	.286	2.48	- .38	.352	2.12	- .42	.321
superb	49.0	2.24	- .13	.357	2.88	.00	.420	2.54	- .07	.393
skirmish	48.5	1.62	- .61	.259	2.54	- .36	.505	2.05	- .49	.412
cloister	47.8	1.39	- .78	.265	2.10	- .80	.600	1.72	- .79	.453
dissent	46.0	2.03	- .29	.389	3.08	.10	.457	2.52	- .11	.465
thud	46.0	1.62	- .54	.316	2.46	- .44	.419	2.01	- .49	.371
straggle	44.8	1.62	- .53	.390	1.99	- .82	.906	1.79	- .66	.695
vicar	44.8	.96	-1.13	.351	1.99	- .84	.445	1.44	-1.00	.423
shank	43.0	1.21	- .86	.340	2.02	- .72	.521	1.59	- .80	.439
ignite	43.0	1.69	- .54	.226	2.69	- .20	.358	2.15	- .38	.340
cryptic	40.0	1.41	- .79	.205	2.23	- .71	.440	1.79	- .75	.338
veterinary	40.0	1.40	- .72	.384	2.18	- .66	.471	1.76	- .79	.428
modulate	39.8	1.32	- .85	.358	2.18	- .69	.724	1.72	- .78	.564
drivel	39.8	1.32	- .82	.423	1.05	-1.81	.540	1.19	-1.28	.689
abduct	38.9	1.51	- .69	.261	2.38	- .45	.556	1.91	- .58	.441
torpor	38.9	1.21	- .95	.311	1.12	-1.68	.849	1.17	-1.29	.720
dill	38.2	1.11	- .98	.681	2.23	- .60	.389	1.63	- .80	.596
ocular	37.9	.90	-1.23	.473	1.84	-1.06	.588	1.34	-1.15	.536
pachyderm	35.9	.37	-1.68	.462	1.42	-1.50	.968	.86	-1.60	.747
spicule	35.9	-.08	-2.01	.513	.76	-2.08	.472	.31	-2.04	.496
grout	35.2	-.03	-2.04	.616	1.14	-1.71	.802	.51	-1.89	.727
echidna	33.5	.01	-2.05	.719	.53	-2.32	.573	.25	-2.18	.669

Table 2

Reliabilities, Standard Errors of Measurement, and

Confidence Limits for Individual and Averaged Estimates

Individual z-Scores for Words	Sample 1 (Lexicographers) (N = 15)	Sample 2 (Other adults) (N = 13)	Combined Samples (N = 28)
Reliability	.869	.746	.801
Mean square error	.134	.259	.203
$S.E._{measmt}$ (z-scores)	.366	.509	.450
$S.E._{measmt}$ (in SFI units)	5.079	7.068	6.253
95% Confidence limits (SFI units)	±9.955	±13.852	±12.256
Averaged z-Scores			
Reliability	.990	.974	.991
$\sigma_{\bar{z}}$	.937	.875	.899
$S.E._{measmt}$	.098	.140	.084
$S.E._{measmt}$ (in SFI units)	1.390	2.224	1.303
95% Confidence limits (SFI units)	±2.724	±4.359	±2.555

Table 3

Mean Standard Deviations of z-Scores, by Ranges of SFI

SFI	No. of Words	Mean SFI	Mean σ of z-scores	
			Sample 1	Sample 2
80-89.9	4	84.7	.365	.334
70-79.9	6	74.5	.323	.426
60-69.9	14	64.1	.319	.348
50-59.9	14	55.5	.255	.454
40-49.9	12	45.1	.314	.491
30-39.9	10	37.4	.482	.646